

Building a discipline-specific aggregate for computing and library and information science

Thomas Krichel

Long Island University, NY, USA

2004-04-13

before I start...

- Thanks to
 - the organizers for inviting me to speak “here”
 - the US Immigration Services and the Department of State for making it impossible to travel
- Apologies for
 - talk being potentially offensive and overly long
 - I will take no offense if you leave the room!
 - not going into much technical details
 - collaboration welcome
 - you can use phone line after the talk...

my view on institutional archives

- They will work a lot better if they are backed-up by discipline-specific aggregation systems.
- Such systems start as basic abstracting and indexing services.
- They evolve into evaluation system that show the scholars relative impact within a neighborhood of other scholars.
- “Such systems are a pie in the sky!”

my beliefs

- Scholarly communication is author-driven.
- Authors act in communities called disciplines.
- In order to change scholarly communication you have simultaneously affect the individual scholar and the discipline.

except for RePEc

- It goes back to efforts I started in 1993 to improve the departmental self-archiving in economics.
- It has grown to a very large *relational* dataset that links
 - document
 - collections of documents
 - authors
 - institutions
- It has achieved a critical mass of data across economics.
- It is slowly getting involved into evaluative work.

recently I have become “reckless”

- rclis, stands for research in computing and library & information science
- Some of my partners in crime are in attendance
 - José Manuel Barrueco Cruz
 - Imma Subirats Coll
 - Antonella De Robbio
- rclis does the same thing as RePEc, but with more modern technology.
- We want to enhance existing and or historical practice, rather than replace it.

historical practice I

- NCSTRL
 - organize the departmental servers of tech reports
 - closed for a while when no funding was available
 - historic data now at <http://www.ncstrl.org>
 - where is the “full” rfc1824 dataset?
- CORR
 - an attempt to design a hybrid between arXiv.org and NCSTRL.
 - has had small numbers of uploads.

historical practice II

- CiteSeer is a pioneering automated citation index
 - 600k documents claimed
 - core collection in computer science but operates beyond
 - entirely automated
- DBLP
 - 450k+ title and collection data, no full text
 - covers conference paper (2/3) and journal papers (1/3)
 - maintained manually

historical practice III

- It is the rest
 - Almost every computer scientist has a homepage.
 - If she is active in research, she will demonstrate that by putting up a few papers.
 - Most of them are not otherwise formally archived.
 - No way to tell what is a paper or what is not.

konz project

- DBLP leads bit of a Cinderella life.
- But it is the crucial component. It has fairly comprehensive coverage of computing as a field. Up to us to find them on the Web.
- This is what the konz project attempts.
 - take paper descriptions from DBLP
 - try to find if they are available for free download on the Web.

aims

- Find out how many papers are freely available.
- Examine the availability of papers as a function of some observable variables.
- Enhance the visibility of these papers by making them available in rclis data portals, to be built.

implementation limitations

- Currently I look at partial subset of DBLP, journal data only, 30k records.
- I only use the title to look for the paper.
- I ignore short titles < 5 words, but no sophisticated way to weed bad titles.
- I only consider full text in Adobe or Microsoft formats.
- I use the Google SOAP API.

implementation details

- At the moment 3,000 lines of Perl and XML code.
- 7 stages of looking at different aspects of the process.
- Software works on a principle of perpetual renewal, i.e. treating a random subset at every
 - good for a development
 - poor to nail down strong statistics

some results

- I can find about 25% of the papers.
- If technically, the software would be better, my guess is I can find 35%
- When I study conference papers I expect better results.
- OAI archives and open access journals are (almost) nowhere to be seen.
- Most CiteSeer links go to references, it does have few full texts in it cache.

if I overcome the limitations

- Give me a bibliographic citation, and konz will fetch it from anywhere on the Internet, not in real time of course.
- No need for formal archiving!
- No need for open access journals, a web version of an eprint will do!
- I expect a reaction to these statements:
- **Crucifixion!**

where is the archive?

- In a bibliography + WWW + konz scheme there is no archive
- Things can disappear at any time,
- so we need a clever scheme to (re)introduce archiving
- rclis does take a cache of the paper, but that is really... reckless

reverse value chain

- Value chain
 - author deposits a preprint
 - get it peer reviewed
 - published in a toll-gated journal/conference proceeding
 - eprint disappears
- Reverse value chain
 - author sends paper to a journal/conference
 - journal/conference says paper has been accepted
 - author is *allowed* to submit a version to an archive

vanity of vanities

- If you open an archive, you ask people to submit, they will not do it!
- If you open an archive where people can only submit by virtue of an especial grace or recognition, they will want to submit.
- There is evidence to that from the RePEc project.
- Now this is a whole other story, on which I have to be brief.

RePEc author service

- It allows authors to associate themselves with the bibliographic data in RePEc.
- These records are used to built an on-line CV, i.e. an evaluative record.
- There is evidence of strong demand from authors to upload papers
 - new papers that they have authored
 - free online versions of already published papers
- It is the personal registration that drives the uploading process, rather than the opposite!

ACIS

- OSI have funded a rewrite of the RePEc author registration system.
- The new software system (ACIS) will have enhanced functionalities
 - allow to associate with citation data
 - allow for uploads of papers
 - calculation of evaluation data for authors
- Project moves slowly but will be done in full. See <http://acis.openlib.org>

conclusion

- Scholarly communication is author driven.
- Authors act in communities called disciplines.
- In order to change scholarly communication you have simultaneously affect the individual scholar and the discipline.
- We can huddle together some document data.
- The crucial part in the personal data.
- We need to work with the living (people) rather than the dead (documents).
- This is what the ACIS project is about.

Thank you for your attention!

<http://openlib.org/home/krichel>